Infrastructure Deficiency Correlation with Pedestrian/Cyclist Crashes: A Data-Driven Approach

Farbod Tavakkoli Khomeini
*Southern Methodist University*, ftavakkolikhomeini@smu.edu

Barbara Minsker
minsken@mail.smu.edu

Michael Hahler
mhaehsler@lyle.smu.edu

Janille Smith-Colin
jsmithcolin@mail.smu.edu

Xinlei Wang
swang@mail.smu.edu

Follow this and additional works at: https://scholar.smu.edu/engineering_civil_etds

Part of the Transportation Engineering Commons

Recommended Citation
Tavakkoli Khomeini, Farbod; Minsker, Barbara; Hahler, Michael; Smith-Colin, Janille; and Wang, Xinlei, "Infrastructure Deficiency Correlation with Pedestrian/Cyclist Crashes: A Data-Driven Approach" (2021). *Civil and Environmental Engineering Theses and Dissertations*. 14.
https://scholar.smu.edu/engineering_civil_etds/14

This Thesis is brought to you for free and open access by the Civil Engineering and Environmental Engineering at SMU Scholar. It has been accepted for inclusion in Civil and Environmental Engineering Theses and Dissertations by an authorized administrator of SMU Scholar. For more information, please visit http://digitalrepository.smu.edu.
Infrastructure Deficiency Correlation with Pedestrian/Cyclist Crashes:
A Data-Driven Approach

Approved by:

Prof. Barbara Minsker
Professor of CEE

Prof. Michael Hahsler
Clinical Associate Professor of CS & OREM

Prof. Janille Smith-Colin
Assistant Professor of CEE

Prof. Xinlei Wang
Professor of STAT
Infrastructure Deficiency Correlation with Pedestrian/Cyclist Crashes: A Data-Driven Approach

A Thesis Presented to the Graduate Faculty of the

Bobby B. Lyle School of Engineering

Southern Methodist University

in

Partial Fulfillment of the Requirements

for the degree of

Master of Science in Civil Engineering

by

Farbod Tavakkoli Khomeini

B.S., Civil Engineering, Sharif University of Technology
Southern Methodist University

August 4, 2021
ACKNOWLEDGMENTS

This work is possible largely by the guidance of my advisor, Prof. Barbara Minsker, along with Prof. Janille Smith-Colin, in the Department of Civil and Environmental, Prof. Xinlei Sherry Wang in the Department of Statistical Science, and Prof. Michael Hahsler in the Department of Computer Science here at Southern Methodist University. I am also grateful to my family, especially my wife, Brittney Gross, for being so patient, supportive, and encouraging throughout my master's degree. I would also like to thank Dr. Amir Samimi and Zheng Li for their help during my research.
A model for identifying the impacts of infrastructure deficiency on road traffic safety is needed to help governments prioritize strategic investments to increase public safety. During the 2010-2019 period, U.S. pedestrian and cyclist fatalities rose by 44% and 36%, respectively, and previous studies have shown a positive correlation between pedestrian/cyclist crashes and low-income areas. To our knowledge, this study is the first to investigate further the reasons behind the higher probability of pedestrian and cyclist crashes in low-income areas. The proposed hypothesis is that the higher probability of pedestrian and cyclist crashes in low-income areas correlates with higher infrastructure deficiencies such as sidewalk, crosswalk, and pavement deficiencies. Ordered logistic regression and K-means clustering techniques have been used in this study to model the impacts of infrastructure deficiency on pedestrian-vehicle and cyclist-vehicle crash frequency at intersections in Dallas, Texas as a case study. The results show that for intersections in low-income areas, the odds of having pedestrian and cyclist crashes are 22% and 34% higher than intersections in middle-income and high-income areas, respectively. For intersections with sidewalk, crosswalk, or pavement deficiencies, the odds of having pedestrian and cyclist crashes are 86%, 15%, and 29% higher, respectively, than intersections without such deficiencies. For intersections with one, two, or three infrastructure deficiencies, the odds of having pedestrian and cyclist crashes are 2.8, 3.0, and 3.2 times higher, respectively, than intersections without infrastructure deficiencies.
# TABLE OF CONTENTS

LIST OF FIGURES ........................................................................................................ viii
LIST OF TABLES ............................................................................................................. ix

Chapter 1 - Introduction .......................................................................................... 1

Chapter 2 - Case Study .......................................................................................... 4

Chapter 3 - Methodology ....................................................................................... 5
  3.1 Research Design Summary ............................................................................. 5
  3.2 Data Sources .................................................................................................... 6
  3.3 Ordered Logistic Regression .......................................................................... 9
  3.4 K-means Clustering ....................................................................................... 10
  3.5 Infrastructure Deficiency Score ..................................................................... 11

Chapter 4 - Results ................................................................................................. 12
  4.1 Ordered Logistic Regression ......................................................................... 12
  4.2 K-means Clustering ....................................................................................... 21

Chapter 5 - Discussion ........................................................................................... 25

Chapter 6 - Conclusions ......................................................................................... 32

Appendix ...................................................................................................................... 34
  3.2.1 Intersection Dataset ................................................................................. 34
  3.2.2 Pedestrian and Cyclist Crash Dataset ...................................................... 35
  3.2.3 Median Household Income by Block Group Dataset ................................. 36
  3.2.4 Pedestrian Trip Generation Dataset .......................................................... 37
  3.2.5 Bus Stop Ridership Data Dataset .............................................................. 38
  3.2.6 Annual Average Daily Traffic Dataset .................................................... 39
  3.2.7 Vulnerability and Disability Dataset ......................................................... 40
  3.2.8 Lane Width Dataset ................................................................................. 42
  3.2.9 No Car Ownership Dataset ....................................................................... 43
  3.2.10 Land Use Dataset .................................................................................... 44
  3.2.11 Sidewalk Dataset .................................................................................... 45
  3.2.12 Crosswalk Dataset .................................................................................. 46
  3.2.13 Pavement Dataset ................................................................................... 48
  3.2.14 Database Validation ............................................................................... 49
  4.2.1 The Box Plot Comparison Among Clusters for Other Features ................. 51
LIST OF FIGURES

Figure 1 - Different phases of this study ................................................................. 5
Figure 2 - Coefficient value comparison of the ordered logistic regression model .......... 14
Figure 3 - Crash probability plot for intersections without infrastructure deficiency ...... 18
Figure 4 - Crash probability plot for intersections with one infrastructure deficiency .... 19
Figure 5 - Crash probability plot for intersections with two infrastructure deficiencies ... 20
Figure 6 - Crash probability plot for intersections with three infrastructure deficiencies ... 21
Figure 7 - Dallas intersections grouped into four clusters ....................................... 23
Figure 8 - Pedestrian and cyclist crash distribution among different clusters .......... 24
Figure 9 - Infrastructure deficiency score distribution among different clusters ...... 24
Figure 10 - Different categories of Dallas intersections ......................................... 26
Figure 11 - An example of intersections with infrastructure deficiency and crashes ...... 26
Figure 12 - An example of intersections with infrastructure deficiency but without crashes .... 27
Figure 13 - An example of intersections without infrastructure deficiency or crashes ...... 27
Figure 14 - An example of intersections without infrastructure deficiency but with crashes .... 28
Figure 15 - Risk analysis ...................................................................................... 28
Figure 16 - Dallas intersections ............................................................................ 34
Figure 17 - Pedestrian and cyclist crashes in Dallas ............................................. 35
Figure 18 - The income level of block groups in Dallas ........................................ 36
Figure 19 - The ITE pedestrian trip generation estimation for liquor store ............... 37
Figure 20 - Bus stops in Dallas ............................................................................. 38
Figure 21 - Annual average daily traffic for different roads/streets in Dallas ............. 39
Figure 22 - Total disability distribution in Dallas .................................................. 40
Figure 23 - Total vulnerability distribution in Dallas ............................................. 41
Figure 24 - The lane width points of streets in Dallas .......................................... 42
Figure 25 - No car ownership distribution in Dallas ............................................. 43
Figure 26 - Land use groups and subsets ............................................................. 44
Figure 27 - A damaged & leave-out sidewalk condition ...................................... 45
Figure 28 - Sidewalk deficiencies in Dallas ......................................................... 46
Figure 29 - An example of a crosswalk deficiency .............................................. 47
Figure 30 - Locations with crosswalk deficiencies in Dallas ............................... 47
Figure 31 - A damaged & leave-out pavement condition ..................................... 48
Figure 32 - Pavement deficiencies in Dallas ......................................................... 49
Figure 33 - An example of errors found in the database ....................................... 50
Figure 34 - AADT distribution among different clusters ....................................... 51
Figure 35 - Bus stop ridership distribution among different clusters ..................... 52
Figure 36 - No car ownership distribution among different clusters ..................... 52
Figure 37 - Pedestrian trip distribution among different clusters ......................... 53
Figure 38 - Vulnerability distribution among different clusters ......................... 53
Figure 39 - Disability distribution among different clusters ............................... 54
LIST OF TABLES

Table 1 - Descriptions of the datasets considered in this study ........................................ 6
Table 2 - A statistical summary of the variables used in this study .................................... 8
Table 3 - Estimates of ordered logistic regression ................................................................. 12
Table 4 - Odds ratio of each independent variable ............................................................... 14
Table 5 - Estimates of the ordered logistic regression under the second scenario ............. 15
Table 6 - Odds ratio of each independent variable under the second scenario .................. 17
Table 7 - Crash costs for highway safety analysis ............................................................... 29
Table 8 - Costs for pedestrian and cyclist infrastructure improvements .......................... 30
Chapter 1 - Introduction

Developing a model for identifying the impacts of infrastructure deficiency on road traffic safety is an important step in helping governments prioritize strategic investments to increase public safety. For instance, the ASCE report card (2021) stated that 43% of the road infrastructure system is deficient (poor or mediocre), and over 36,000 people die on the nation's roads every year. The ASCE report (2021) emphasized that “Federal, state, and local governments will need to prioritize strategic investments dedicated to improving and preserving roadway conditions that increase public safety.” In addition, according to the World Health Organization (2018), more than half of the deaths in road traffic crashes every year around the world are among vulnerable road users, including pedestrians and cyclists. In recent years, pedestrian and cyclist fatalities in the U.S. have risen at an unprecedented rate. During the 2010-2019 period, U.S. pedestrian and cyclist fatalities rose by 44% and 36%, respectively, while all other traffic fatalities combined rose by less than 5% during this same period (NHTSA, 2020).

A few previous studies have investigated the correlation between pedestrian/cyclist crashes and low-income areas. For instance, Loukaitou-Sideris and Liggett (2007) found that “pedestrian accidents are more likely to occur in low-income, minority neighborhoods once other aspects of risk are controlled for.” Laflamme and Diderichsen (2000) stated that “for most types of traffic injuries, mortality and morbidity are often higher among children from lower social positions and in more deprived socio-economic areas.” To our knowledge, this study is the first to further investigate the reasons behind the higher probability of pedestrian and cyclist crashes in low-income areas. The proposed hypothesis is that the higher probability of pedestrian and cyclist
crashes in low-income areas correlates with higher infrastructure deficiencies such as sidewalk, crosswalk, and pavement deficiencies.

Various modeling techniques have been proposed previously to model crash frequency. Cottrill and Thakuriah (2010) used Poisson regression to model crash frequency in Chicago, Illinois, with road length, suitability for walking, transit availability, crime rates, income, and presence of children as the feature variables. Hess et al. (2004) used logistic regression to model crash frequency in Washington State, with vehicle volume, the number of traffic lanes, transit stop usage, and retail location size as the feature variables. Pulugurtha and Sambhara (2011) used negative binomial regression to model crash frequency in Charlotte, North California, with pedestrian volume, vehicle volume, bus stop number, land use, and population as the feature variables.

This study uses ordered logistic regression and K-means clustering techniques to model the impacts of infrastructure deficiency on pedestrian-vehicle and cyclist-vehicle crash frequency at intersections, considering the following feature variables: sidewalk deficiency, crosswalk deficiency, pavement deficiency, income level, bus stop ridership, pedestrian trip generation, vulnerability, car ownership, disability, number of bars, average annual daily traffic, and lane width. The performance of the ordered logistics regression model is compared with the K-mean clustering model, and both models are used to explore how infrastructure deficiencies affect pedestrian and cyclist crash frequency by income levels. A case study in Dallas, Texas, is examined to develop and test the models.

This thesis is organized into six chapters. The first chapter introduces the objectives and scope of the research. The case study in Dallas, Texas, is described in the second chapter. The third chapter discusses the methodology of this study, including the data resources, research design
summary, and overview of the ordered logistic regression and K-means clustering models. The fourth chapter gives the results of the study, followed by further discussion in the next chapter. Finally, the last chapter provides conclusions, limitations, and recommendations for further research.
Chapter 2 - Case Study

This study examines a case study in Dallas, Texas, which is the third-largest city in Texas and the ninth most populated city in the U.S. (Aman, J. J. C., & Smith-Colin, J., 2020; Census, 2021). The city has seen a 37% rise in pedestrian and cyclist crashes (TxDOT, 2021) and an 8.4% increase in population (Census, 2019) during the 2010-2019 period. Dallas is also ranked as the 8th highest city for income inequality in the United States based on the Gini coefficient (Census, 2019). These factors make Dallas particularly suitable for examining the relationship between infrastructure deficiencies, income, and pedestrian and cyclist crashes.

The geospatial unit of analysis is intersections, which have higher risks due to more conflict points between travelers and vehicles (Ukkusuri, S., Miranda-Moreno, L. F., Ramadurai, G., & Isa-Tavarez, J., 2012). The models for this study are fit using data on pedestrian-vehicle and cyclist-vehicle crashes collected from 2015 to 2019 at the intersection level by the TxDOT Crash Records Information System (CRIS). A total of 3,795 geocoded pedestrian and cyclist crashes occurred in this period at 33,965 intersections. The crash risk (dependent variable) is defined as the number of pedestrian and cyclist crashes at each intersection during this period.
Chapter 3 - Methodology

This chapter describes the methodology used in this study and includes five subsequent sections. Section 1 provides a summary of the four phases of this study. Section 2 describes the datasets and gives a statistical summary of the variables used in this study. Section 3 provides basic concepts of the ordered logistic regression model used in this study. Section 4 summarizes the K-means clustering model procedure. Finally, the last section describes how the infrastructure deficiency score is calculated for each intersection.

3.1 Research Design Summary

Figure 1 provides a summary of the four phases of this study: input dataset, data pre-processing, modeling, and output. In Phase 1, the correlated features with pedestrian-vehicle and cyclist-vehicle crashes at intersections were explored and selected based on the literature review. In Phase 2, data pre-processing steps were taken to prepare a dataset for modeling purposes. In Phase 3, ordered logistic regression and K-means clustering models were created and evaluated. In Phase 4, the results of the models were shown and interpreted.

Figure 1 - Different phases of this study
3.2 Data Sources

This section describes the datasets considered in this study, provides information about each dataset's data source and criteria, and explains why some features are excluded from the analysis of this study. Table 1 summarizes the datasets considered in this study. A more detailed explanation regarding each dataset description and criteria is presented in the Appendix Sections 2.2.1 to 2.2.14.

Table 1 - Descriptions of the datasets considered in this study

<table>
<thead>
<tr>
<th>Dataset Name</th>
<th>Description</th>
<th>Criteria</th>
<th>Source</th>
</tr>
</thead>
</table>
| Crash              | Pedestrian and cyclist crashes at intersections  | 0: Low Risk 
 <= 2: Medium Risk 
 > 2: High Risk                                   | CRIS     |
| Income             | 5-year estimate of median household income by block group | < 33%: Low Income 
 <= 66%: Medium Income 
 > 66%: High Income                                      | Census   |
<p>| Pedestrian Trip    | Number of pedestrian trips generated based on the land use | -                                              | ITE      |
| Bus Stop Ridership | Bus usage at each bus stop                       | -                                              | DART     |
| AADT               | Number of vehicles of a road for a year divide by 365 days | -                                              | TxDOT    |
| Vulnerability      | Number of people under five or older than 64 years old | -                                              | Census   |
| Disability         | Number of people with disabilities               | -                                              | Census   |
| Lane Width         | Lane width of streets                             | -                                              | Dallas EGIS |
| No Car Ownership   | Number of tenures without access to cars          | -                                              | Census   |</p>
<table>
<thead>
<tr>
<th>Dataset Name</th>
<th>Description</th>
<th>Criteria</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Land Use</td>
<td>Human use of land</td>
<td>Seven categories, including residential, commercial, industrial, entertainment, vacant, and educational</td>
<td>Dallas EGIS</td>
</tr>
<tr>
<td>Light Rail Transit Ridership</td>
<td>LRT usage at each stop</td>
<td>-</td>
<td>DART</td>
</tr>
<tr>
<td>Ramp</td>
<td>Location of ramps</td>
<td>No ramp: Deficient</td>
<td>Dallas EGIS</td>
</tr>
<tr>
<td>Traffic Signals</td>
<td>Location of traffic signals</td>
<td>No traffic signal: Deficient</td>
<td>Dallas EGIS</td>
</tr>
<tr>
<td>Race</td>
<td>Race information for each block group</td>
<td>Five categories, including White, Black, Hispanic, American Indian, and Asian</td>
<td>Census</td>
</tr>
<tr>
<td>Education</td>
<td>Education information for each block group</td>
<td>Five categories, including no education, high school, some college, bachelor, and post-grad</td>
<td>Census</td>
</tr>
<tr>
<td>Vehicle Trip Generation</td>
<td>Number of vehicle trips generated based on the land use</td>
<td>-</td>
<td>ITE</td>
</tr>
<tr>
<td>Street Lights</td>
<td>Location of street lights</td>
<td>-</td>
<td>Dallas EGIS</td>
</tr>
<tr>
<td>Traffic Signs</td>
<td>Location of traffic sings</td>
<td>No traffic sign: Deficient</td>
<td>Dallas EGIS</td>
</tr>
<tr>
<td>Park Centroids</td>
<td>Location of parks</td>
<td>-</td>
<td>Dallas EGIS</td>
</tr>
<tr>
<td>Trinity Railway Express Ridership</td>
<td>TRE usage at each railway express stop</td>
<td>-</td>
<td>DART</td>
</tr>
<tr>
<td>Rail Stations</td>
<td>Location of rail stations</td>
<td>-</td>
<td>Dallas EGIS</td>
</tr>
<tr>
<td>Urban Spaces</td>
<td>Contemporary public space classification</td>
<td>Four categories, including negative space, positive space, ambiguous space, and private space</td>
<td>Carmona, M. (2010)</td>
</tr>
<tr>
<td>Trails</td>
<td>Location of trails</td>
<td>-</td>
<td>Dallas EGIS</td>
</tr>
<tr>
<td>Sidewalk</td>
<td>Sidewalk quality of each street</td>
<td>Damaged, none, or leave-out: Deficient</td>
<td>DPW</td>
</tr>
</tbody>
</table>
During the modeling process for this study, many variables were excluded that had a high correlation with other covariates, had a p-value less than the alpha level (0.05), or had incomplete spatial coverage. The variables excluded for collinearity (high correlation) are ramp deficiency, park centroids, rail stations, trails, residential buildings, commercial buildings, entertainment buildings, industrial buildings, medical buildings, vacant, educational buildings, positive space, ambiguous space, negative space, and private space. For statistical insignificance reasons (p-value less than 0.05), the variables excluded are race group, street lights, education level, flood probability, light trail ridership, Trinity Railway express ridership, and vehicle trip generation. Finally, the variables excluded for incomplete spatial coverage are traffic signals, traffic signs, and bike lanes. Table 2 indicates a statistical summary of the final variables used in this study.

Table 2 - A statistical summary of the variables used in this study

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>N. Observations</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Median</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crash</td>
<td>52,011</td>
<td>0.2</td>
<td>0.6</td>
<td>0</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>Income</td>
<td>69,209</td>
<td>69,704</td>
<td>43,238</td>
<td>57,246</td>
<td>9,052</td>
<td>250,000</td>
</tr>
<tr>
<td>Pedestrian Trip</td>
<td>4,731</td>
<td>610</td>
<td>1,032</td>
<td>253</td>
<td>0.2</td>
<td>22,341</td>
</tr>
<tr>
<td>Bus Stop Ridership</td>
<td>14,627</td>
<td>29.3</td>
<td>1,934</td>
<td>0</td>
<td>0</td>
<td>10,349</td>
</tr>
<tr>
<td>AADT</td>
<td>790</td>
<td>9,785</td>
<td>46,176</td>
<td>288</td>
<td>1</td>
<td>1,128,745</td>
</tr>
<tr>
<td>Vulnerability</td>
<td>3,181</td>
<td>932</td>
<td>455</td>
<td>869</td>
<td>0</td>
<td>2,285</td>
</tr>
<tr>
<td>Variable Name</td>
<td>N. Observations</td>
<td>Mean</td>
<td>Standard Deviation</td>
<td>Median</td>
<td>Min</td>
<td>Max</td>
</tr>
<tr>
<td>-------------------</td>
<td>-----------------</td>
<td>------</td>
<td>--------------------</td>
<td>--------</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Disability</td>
<td>790</td>
<td>544</td>
<td>327</td>
<td>501</td>
<td>0</td>
<td>1.851</td>
</tr>
<tr>
<td>Lane Width</td>
<td>790</td>
<td>12</td>
<td>2.3</td>
<td>11.9</td>
<td>0</td>
<td>36.5</td>
</tr>
<tr>
<td>No Car Ownership</td>
<td>790</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>Bar</td>
<td>30,409</td>
<td>0.1</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>Sidewalk</td>
<td>361,843</td>
<td>0.62</td>
<td>0.48</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Crosswalk</td>
<td>84,524</td>
<td>0.081</td>
<td>0.27</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Pavement</td>
<td>2,972</td>
<td>0.48</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

3.3 Ordered Logistic Regression

Ordered/ordinal logistic regression, also known as proportional odds logistic regression, uses log-odds of cumulative probabilities for an ordinal response with the general formula of:

$$\text{Logit}(P(Y \leq j|X)) = \alpha_j + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$  \hspace{1cm} (1)

Where Y is an ordinal outcome with J categories, $X_j$ are the independent variables, $\beta_j$ are the coefficients of independent variables, and $\alpha_j$ are the model's intercept for each category (J).

Observed crash frequencies (Y) near intersections can be defined as:

$Y = 1 \rightarrow \text{Low Risk (no crash)}$

$Y = 2 \rightarrow \text{Medium Risk (one or two crashes)}$  \hspace{1cm} (2)

$Y = 3 \rightarrow \text{High Risk (more than two crashes)}$

Therefore, the probability that an individual intersection belongs to either of the three categories is defined as:
\( P(Y = 1) = \Lambda(\beta X) \)

\( P(Y = 2) = P(Y \leq 2) - P(Y \leq 1) = \Lambda(\alpha_1 + \beta X) - \Lambda(\beta X) \)

\( P(Y = 3) = P(Y \leq 3) - P(Y \leq 2) = \Lambda(\alpha_2 + \beta X) - \Lambda(\alpha_1 + \beta X) \)

(3)

Where \( \Lambda(\cdot) \) is the standard logistic cumulative distribution function, and \( \alpha_1 \) and \( \alpha_2 \) represent the lower and upper thresholds (cutpoints) for the outcome. The odds of an outcome \( i \) may be defined as:

\[
\frac{P(Y = i)}{1 - P(Y = i)} = \exp(\beta_0 + \beta_1 X)
\]

(4)

### 3.4 K-means Clustering

Various studies have proposed unsupervised machine learning algorithms to classify unlabeled micromobility datasets (Aman, J. J., & Smith-Colin, J. , 2021; Aman, J. J., Smith-Colin, J., & Zhang, W., 2021). K-means clustering algorithm is one of the most common and famous unsupervised machine learning techniques. K-means is a partitional or non-hierarchical clustering method partitions a set of data objects into non-overlapping homogeneous clusters such that each data object is in exactly one cluster. Clusters are created based on the closeness of the observations from the clusters’ means as measured by the Euclidean distance function (5).

\[
d_{ij} = \left( \sum_{k=1}^{P} (x_{ik} - x_{jk})^2 \right)^{1/2}
\]

(5)

The K-means clustering algorithm can be simplified and explained in a few steps as:

1. Randomly taking K points as initial centroids (center of clusters)
2. Assigning each point to the nearest centroid
3. Forming the K clusters (each cluster is associated with a centroid)
4. Updating the centroids in each cluster (centroids are the mean of the points in each cluster)

5. Iteratively repeating until centroids are stable.

3.5 Infrastructure Deficiency Score

A more summative variable is included by adding an infrastructure deficiency score to the models that represents how sidewalk, crosswalk, and pavement deficiencies as a whole affect pedestrians and cyclist crashes at intersections. The infrastructure deficiency score is defined as:

\[
\text{Infrastructure Deficiency Score (Z)} = \text{Sidewalk Deficiency (0 or 1)} + \text{Crosswalk Deficiency (0 or 1)} + \text{Pavement Deficiency (0 or 1)}
\]  

Therefore, the infrastructure deficiency score belongs to either of the four categories is defined as:

\[Z = 0 \rightarrow \text{No infrastructure deficiency}\]

\[Z = 1 \rightarrow \text{One infrastructure deficiency}\]

\[Z = 2 \rightarrow \text{Two infrastructure deficiencies}\]

\[Z = 3 \rightarrow \text{Three infrastructure deficiencies}\]
Chapter 4 - Results

This chapter discusses the results of this study and includes two subsequent sections. Section 1 summarizes the ordered logistic regression model results, while Section 2 summarizes the results of the K-means clustering model.

4.1 Ordered Logistic Regression

This section summarizes the results of the ordered logistic regression applied to the 2015-2019 pedestrian and cyclist crash data at intersections in Dallas, Texas. Table 3 gives the modeling estimates (log odds) for the correlation between independent variables and the pedestrian and cyclist crash frequency. Table 3 shows that all of the independent variable estimates are statistically significant because each coefficient p-value is less than 0.05.

Table 3 - Estimates of ordered logistic regression

<table>
<thead>
<tr>
<th>Coefficient Name</th>
<th>Coefficient Value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medium Income</td>
<td>-0.199</td>
<td>1.30E-06</td>
</tr>
<tr>
<td>High Income</td>
<td>-0.294</td>
<td>4.42E-07</td>
</tr>
<tr>
<td>Sidewalk Deficiency</td>
<td>0.623</td>
<td>2.12E-39</td>
</tr>
<tr>
<td>Crosswalk Deficiency</td>
<td>0.136</td>
<td>0.03</td>
</tr>
<tr>
<td>Pavement Deficiency</td>
<td>0.251</td>
<td>3.12E-09</td>
</tr>
<tr>
<td>Pedestrian Trip</td>
<td>0.089</td>
<td>0.039</td>
</tr>
<tr>
<td>Pedestrian Trip^2</td>
<td>-0.01</td>
<td>0.026</td>
</tr>
<tr>
<td>Bus Stop Ridership</td>
<td>0.245</td>
<td>4.90E-156</td>
</tr>
<tr>
<td>AADT</td>
<td>0.316</td>
<td>2.00E-307</td>
</tr>
<tr>
<td>Lane Width</td>
<td>-0.366</td>
<td>0.001</td>
</tr>
<tr>
<td>No Car Ownership</td>
<td>1.246</td>
<td>2.21E-06</td>
</tr>
<tr>
<td>Bar</td>
<td>0.531</td>
<td>1.37E-18</td>
</tr>
<tr>
<td>Coefficient Name</td>
<td>Coefficient Value</td>
<td>p-value</td>
</tr>
<tr>
<td>------------------</td>
<td>-------------------</td>
<td>---------</td>
</tr>
<tr>
<td>Vulnerability</td>
<td>-0.202</td>
<td>1.31E-06</td>
</tr>
<tr>
<td>Disability</td>
<td>0.474</td>
<td>0.002</td>
</tr>
<tr>
<td>(Disability)^2</td>
<td>-0.044</td>
<td>0.001</td>
</tr>
</tbody>
</table>

Regarding the estimates in Table 3, the variables *medium-income, high-income, lane width, and vulnerability* are negatively associated with pedestrian and cyclist crash frequency at intersections. In contrast, *sidewalk deficiency, crosswalk deficiency, pavement deficiency, bus stop ridership, AADT, no car ownership, and the number of bars* have positive correlations with pedestrian and cyclist crash frequency at intersections. *Pedestrian trips and disability* features had a non-linear relationship with the pedestrian and cyclist crash frequency at intersections. That is why the second power of *pedestrian trips and disability* features is added to the model. The first powers of both *pedestrian trips and disability* are positively associated with pedestrian and cyclist crash frequency at intersections, while their second power is negatively correlated.

Figure 2 visualizes a comparison among coefficient values of the ordered logistic regression model, indicating whether each coefficient is negatively or positively associated with pedestrian and cyclist crash frequency at intersections. Figure 2 shows that *no car ownership* has the highest positive correlation with pedestrian and cyclist crash frequency at intersections, followed by *sidewalk deficiency, bars, disability, AADT, pavement deficiency, bus stop ridership, crosswalk deficiency, and pedestrian trips*. For example, this means that when *car ownership* is decreased, then crashes are most likely to be higher. In contrast, the *lane width* has the highest negative correlation with pedestrian and cyclist crash frequency at intersections, followed by *high-income, vulnerability, and medium-income.*
Figure 2 - Coefficient value comparison of the ordered logistic regression model

Table 4 presents the odds ratio (OR) for each independent variable, calculated by Equation 4. Table 4 also shows that, for each variable, the 97.5% confidence interval does not cross one, and hence the odds ratios are statistically significant. The results show that for intersections in low-income areas, the odds of having more pedestrian and cyclist crashes are 22% and 34% higher than intersections in medium-income and high-income areas, respectively. For intersections with sidewalk, crosswalk, or pavement deficiencies, the odds of having more pedestrian and cyclist crashes are 86%, 15%, and 29% higher than intersections without sidewalk, crosswalk, or pavement deficiencies, respectively.

Table 4 - Odds ratio of each independent variable

<table>
<thead>
<tr>
<th>Variable name</th>
<th>Odds Ratio</th>
<th>CI 2.5%</th>
<th>CI 97.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medium Income</td>
<td>0.82</td>
<td>0.74</td>
<td>0.908</td>
</tr>
<tr>
<td>High Income</td>
<td>0.745</td>
<td>0.665</td>
<td>0.835</td>
</tr>
<tr>
<td>Sidewalk Def</td>
<td>1.864</td>
<td>1.7</td>
<td>2.047</td>
</tr>
<tr>
<td>Variable name</td>
<td>Odds Ratio</td>
<td>CI 2.5%</td>
<td>CI 97.5%</td>
</tr>
<tr>
<td>-------------------------</td>
<td>------------</td>
<td>---------</td>
<td>----------</td>
</tr>
<tr>
<td>Crosswalk Def</td>
<td>1.146</td>
<td>1.012</td>
<td>1.295</td>
</tr>
<tr>
<td>Pavement Def</td>
<td>1.286</td>
<td>1.183</td>
<td>1.397</td>
</tr>
<tr>
<td>Pedestrian Trip</td>
<td>1.093</td>
<td>1.005</td>
<td>1.19</td>
</tr>
<tr>
<td>Pedestrian Trip^2</td>
<td>0.99</td>
<td>0.981</td>
<td>0.999</td>
</tr>
<tr>
<td>Bus Stop Ridership</td>
<td>1.278</td>
<td>1.255</td>
<td>1.302</td>
</tr>
<tr>
<td>AADT</td>
<td>1.372</td>
<td>1.35</td>
<td>1.395</td>
</tr>
<tr>
<td>Lane Width</td>
<td>0.694</td>
<td>0.56</td>
<td>0.861</td>
</tr>
<tr>
<td>No Car Ownership</td>
<td>3.477</td>
<td>1.79</td>
<td>6.72</td>
</tr>
<tr>
<td>Bar</td>
<td>1.701</td>
<td>1.51</td>
<td>1.913</td>
</tr>
<tr>
<td>Vulnerability</td>
<td>0.817</td>
<td>0.737</td>
<td>0.906</td>
</tr>
<tr>
<td>Disability</td>
<td>1.607</td>
<td>1.196</td>
<td>2.211</td>
</tr>
<tr>
<td>(Disability)^2</td>
<td>0.957</td>
<td>0.932</td>
<td>0.981</td>
</tr>
</tbody>
</table>

To further investigate these findings, Table 5 shows the modeling estimates for a second scenario adding the overall infrastructure deficiency score defined in Equation (6) to the model. The coefficient values remain nearly the same, which means the relationship between independent variables and pedestrian and cyclist crash frequency explained in Table 3 is similar for overall infrastructure deficiency. This means that the individual deficiency variables are sufficient, and the overall variable is unnecessary.

Table 5 - Estimates of the ordered logistic regression under the second scenario

<table>
<thead>
<tr>
<th>Coefficient Name</th>
<th>Coefficient Value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medium Income</td>
<td>-0.203</td>
<td>9.58E-05</td>
</tr>
<tr>
<td>High Income</td>
<td>-0.306</td>
<td>1.49E-07</td>
</tr>
<tr>
<td>One Infrastructure Deficiency</td>
<td>1.039</td>
<td>1.34E-58</td>
</tr>
</tbody>
</table>
Table 6 presents the odds ratio (OR) for each independent variable under the second scenario with overall infrastructure deficiency. Table 6 shows that all of the independent variable estimates are still statistically significant, and the odds ratios have remained almost the same. The results show that for intersections with one, two, or three infrastructure deficiencies, the odds of having more pedestrian and cyclist crashes are 2.8, 3.0, and 3.2 times higher than intersections without infrastructure deficiencies, respectively. According to Zheng (2021), low-income areas in Dallas are five times more likely to have highly deficient infrastructure than high-income areas. Table 6 indicates that the higher the infrastructure deficiency, the higher the odds of having pedestrian and cyclist crashes, which supports the hypothesis of this study that the higher probability of pedestrian and cyclist crashes in low-income areas correlates with higher infrastructure deficiencies.
Table 6 - Odds ratio of each independent variable under the second scenario

<table>
<thead>
<tr>
<th>Coefficient Name</th>
<th>Odds Ratio</th>
<th>CI 2.5%</th>
<th>CI 97.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medium Income</td>
<td>0.816</td>
<td>0.737</td>
<td>0.904</td>
</tr>
<tr>
<td>High Income</td>
<td>0.736</td>
<td>0.656</td>
<td>0.825</td>
</tr>
<tr>
<td>One Infrastructure Deficiency</td>
<td>2.827</td>
<td>2.494</td>
<td>3.21</td>
</tr>
<tr>
<td>Two Infrastructure Deficiencies</td>
<td>2.992</td>
<td>2.63</td>
<td>3.41</td>
</tr>
<tr>
<td>Three Infrastructure Deficiencies</td>
<td>3.204</td>
<td>2.608</td>
<td>3.924</td>
</tr>
<tr>
<td>Bus Stop Ridership</td>
<td>1.275</td>
<td>1.252</td>
<td>1.298</td>
</tr>
<tr>
<td>AADT</td>
<td>1.384</td>
<td>1.361</td>
<td>1.407</td>
</tr>
<tr>
<td>Lane Width</td>
<td>0.719</td>
<td>0.581</td>
<td>0.891</td>
</tr>
<tr>
<td>No Car Ownership</td>
<td>3.803</td>
<td>1.96</td>
<td>7.345</td>
</tr>
<tr>
<td>Bar</td>
<td>1.707</td>
<td>1.515</td>
<td>1.921</td>
</tr>
<tr>
<td>Vulnerability</td>
<td>0.817</td>
<td>0.737</td>
<td>0.906</td>
</tr>
</tbody>
</table>

Considering the second scenario’s results for the ordered logistic regression model, the probability of pedestrian and cyclist crashes at intersections are calculated based on their income level, risk level, and infrastructure deficiency score. In this calculation, the values of other covariates are fixed at their mean levels, and only income level, risk level, and infrastructure deficiency score covariates are changed.

Figure 3 shows that in medium- and high-risk intersections without infrastructure deficiency, the crash probability is 29% and 35% higher in low-income areas compared to high-income areas, respectively. In contrast, in low-risk intersections without infrastructure deficiency, the probability of not having crashes is 4.7% lower in low-income areas compared to high-income areas. Figure 4 shows that in medium-risk and high-risk intersections with one infrastructure deficiency, the crash probability is 21% and 34% higher, respectively, in low-income areas.
compared to high-income areas. In contrast, for low-risk intersections with one infrastructure deficiency, the probability of not having crashes is 11% lower in low-income areas compared to high-income areas.

Figure 3 - Crash probability plot for intersections without infrastructure deficiency
Figure 5 shows that in medium-risk and high-risk intersections with two infrastructure deficiencies, the crash probability is 20% and 34% higher, respectively, in low-income areas compared to high-income areas. In contrast, in low-risk intersections with two infrastructure deficiencies, the probability of not having crashes is 11.2% lower in low-income areas compared to high-income areas.
Finally, Figure 6 shows that in medium-risk and high-risk intersections with three infrastructure deficiencies, the crash probability is 20% and 34% higher, respectively, in low-income areas compared to high-income areas. In contrast, in low-risk intersections with three infrastructure deficiencies, the probability of not having crashes is 11.7% lower in low-income areas compared to high-income areas.
4.2 K-means Clustering

This section summarizes the K-means clustering results applied to the 2015-2019 pedestrian and cyclist crash data at intersections in Dallas, Texas. The algorithm has grouped the data into four clusters based on their characteristics. Figure 7 visualizes the clustering results for all Dallas intersections. Figure 8 and Figure 9 give box plots for comparing pedestrian and cyclist crash and infrastructure deficiency score distributions among clusters. The box plot comparison among clusters for other features is presented in Appendix Section 4.2.1. Figure 8 and Figure 9 include information about the Anova test p-value (red text), the t-test p-value between every set of
two groups (black text on brackets), the mean of the y-axis in each group (black points in the middle of each box), and the mean of the y-axis among all groups (purple dash line).

Figure 8 shows that intersections grouped in Cluster Three (red points in Figure 7) have significantly higher pedestrian and cyclist crash rates compared to intersections grouped in other clusters. Furthermore, Figure 9 shows that those intersections grouped in Cluster Three also have the highest infrastructure deficiency score among other clusters, signifying that the higher pedestrian and cyclist crash rates correlate with higher infrastructure deficiency in Cluster Three. In addition, 53% of the intersections in Cluster Three are located in low-income areas, 24% are located in medium-income areas, and 24% are located in high-income areas. Therefore, a higher proportion of the intersections in Cluster Three are located in low-income areas and have a higher pedestrian and cyclist crash rate than other clusters and a higher infrastructure deficiency score. These findings support the hypothesis of this study and confirm the ordered logistic regression model results.
Figure 7 - Dallas intersections grouped into four clusters
Figure 8 - Pedestrian and cyclist crash distribution among different clusters

Figure 9 - Infrastructure deficiency score distribution among different clusters

Anova, $p < 2.2e-16$

Anova, $p = 4.8e-11$
Chapter 5 - Discussion

The intersections in Dallas can be categorized into four different groups, shown in Figure 12, based on their crash frequency and infrastructure deficiency level. Figure 10 indicates that those intersections with infrastructure deficiencies and crashes (red group) are associated with high annual average daily traffic (AADT), bus stop ridership, and pedestrian trips. Figure 11 provides an example of an intersection in the red group, which has sidewalk, crosswalk, and pavement deficiencies and three crashes. Many travelers and vehicles visit this intersection, and infrastructure deficiencies are high, potentially contributing to delete more crashes.

Conversely, intersections with or without infrastructure deficiency and with no crashes (green and blue groups) are associated with low AADT, bus stop ridership, and pedestrian trips. Fewer travelers and vehicles visit these intersections, resulting in fewer crashes. Figure 12 and Figure 13 represent two examples of green and blue groups, respectively. Finally, intersections without infrastructure deficiency but with crashes (orange group) are intersections with only one or two crashes (mostly one) during the 2015-2019 period. Figure 14 shows an example of an intersection in the orange group.
Figure 10 - Different categories of Dallas intersections

![Diagram showing categories of Dallas intersections]

- **Intersections**
  - **With Infrastructure Deficiency**
    - With Crash
    - Without Crash
  - **Without Infrastructure Deficiency**
    - With Crash
    - Without Crash

Figure 11 - An example of intersections with infrastructure deficiency and crashes

![Image of an intersection with infrastructure deficiency and crashes]
Figure 12 - An example of intersections with infrastructure deficiency but without crashes

Figure 13 - An example of intersections without infrastructure deficiency or crashes
Figure 14 - An example of intersections without infrastructure deficiency but with crashes

The results of this study can be summarized with the risk analysis diagram shown in Figure 15. The figure illustrates that high pedestrian and cyclist risk is associated with both high infrastructure deficiency and low-income areas. On the other hand, low pedestrian and cyclist risk is correlated with both low infrastructure deficiency and high-income areas.

Figure 15 - Risk analysis
In terms of policy implications of these findings, Table 7 (MASS, 2019) and Table 8 (Bushell, M. A., Poole, B. W., Zegeer, C. V., & Rodriguez, D. A., 2013) show that the economic costs of crash injuries to society are much higher than the economic costs of infrastructure maintenance and improvement. Given the significantly higher rates of pedestrian and cyclist crashes at intersections with infrastructure deficiencies, this study suggests that increased investments in sidewalks, pavements, and crosswalks may be worthwhile. Furthermore, these investments should prioritize low-income areas with high annual average daily traffic (AADT), bus stop ridership, and pedestrian trips. Note, however, that correlations are not necessarily causation, and further investigation is needed to confirm these findings. A before-and-after study is highly suggested for those locations where infrastructure maintenance and improvements have been made. This would help to confirm whether there is a causal relationship between infrastructure deficiency and pedestrian and cyclist crash frequency.

Table 7 - Crash costs for highway safety analysis

<table>
<thead>
<tr>
<th>Crash Severity</th>
<th>Crash Severity Defined</th>
<th>2019 Recommended Comprehensive Crash Unit Costs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>K</strong></td>
<td>Crashes involving a Fatal Injury</td>
<td>$16,257,800</td>
</tr>
<tr>
<td><strong>A</strong></td>
<td>Crashes involving a Serious Injury</td>
<td>$941,300</td>
</tr>
<tr>
<td><strong>B</strong></td>
<td>Crashes involving a Non-serious Injury</td>
<td>$284,600</td>
</tr>
<tr>
<td><strong>C</strong></td>
<td>Crashes involving a Possible Injury</td>
<td>$179,600</td>
</tr>
<tr>
<td><strong>O</strong></td>
<td>Crashes involving No Injuries</td>
<td>$16,700</td>
</tr>
<tr>
<td><strong>KA</strong></td>
<td>Crashes involving a Fatal Injury OR a Serious Injury</td>
<td>$2,764,700</td>
</tr>
<tr>
<td><strong>KAB</strong></td>
<td>Crashes involving a Fatal Injury OR a Serious Injury OR a Non-Serious Injury</td>
<td>$706,100</td>
</tr>
</tbody>
</table>
Crash Severity

**Crash Severity Defined**

**2019 Recommended Comprehensive Crash Unit Costs**

<table>
<thead>
<tr>
<th>KABC</th>
<th>Crashes involving a Fatal Injury OR an Injury of any type</th>
<th>$441,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>KABCO</td>
<td>Any crash severity</td>
<td>$121,400</td>
</tr>
</tbody>
</table>

Table 8 - Costs for pedestrian and cyclist infrastructure improvements

<table>
<thead>
<tr>
<th>Infrastructure</th>
<th>Description</th>
<th>Median</th>
<th>Average</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Cost Unit</th>
<th>Number of Sources (Observations)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crosswalk</td>
<td>High Visibility Crosswalk</td>
<td>$3,070</td>
<td>$2,540</td>
<td>$600</td>
<td>$5,710</td>
<td>Each</td>
<td>4 (4)</td>
</tr>
<tr>
<td>Crosswalk</td>
<td>Striped Crosswalk</td>
<td>$340</td>
<td>$770</td>
<td>$110</td>
<td>$2,090</td>
<td>Each</td>
<td>8 (8)</td>
</tr>
<tr>
<td>Crosswalk</td>
<td>Striped Crosswalk</td>
<td>$5.87</td>
<td>$8.51</td>
<td>$1.03</td>
<td>$26</td>
<td>Linear Foot</td>
<td>12 (48)</td>
</tr>
<tr>
<td>Crosswalk</td>
<td>Striped Crosswalk</td>
<td>$6.32</td>
<td>$7.38</td>
<td>$1.06</td>
<td>$31</td>
<td>Square Foot</td>
<td>5 (15)</td>
</tr>
<tr>
<td>Sidewalk</td>
<td>Asphalt Paved Shoulder</td>
<td>$5.81</td>
<td>$5.56</td>
<td>$2.96</td>
<td>$7.65</td>
<td>Square Foot</td>
<td>1 (4)</td>
</tr>
<tr>
<td>Sidewalk</td>
<td>Asphalt Sidewalk</td>
<td>$16</td>
<td>$35</td>
<td>$6.02</td>
<td>$150</td>
<td>Linear Foot</td>
<td>7 (11)</td>
</tr>
<tr>
<td>Sidewalk</td>
<td>Brick Sidewalk</td>
<td>$60</td>
<td>$60</td>
<td>$12</td>
<td>$160</td>
<td>Linear Foot</td>
<td>9 (9)</td>
</tr>
<tr>
<td>Sidewalk</td>
<td>Concrete Paved Shoulder</td>
<td>$6.10</td>
<td>$6.64</td>
<td>$2.79</td>
<td>$58</td>
<td>Square Foot</td>
<td>1 (11)</td>
</tr>
<tr>
<td>Sidewalk</td>
<td>Concrete Sidewalk</td>
<td>$27</td>
<td>$32</td>
<td>$2.09</td>
<td>$410</td>
<td>Linear Foot</td>
<td>46 (164)</td>
</tr>
<tr>
<td>Sidewalk</td>
<td>Concrete Sidewalk - Patterned</td>
<td>$38</td>
<td>$36</td>
<td>$11</td>
<td>$170</td>
<td>Linear Foot</td>
<td>4 (5)</td>
</tr>
<tr>
<td>Sidewalk</td>
<td>Concrete Sidewalk - Stamped</td>
<td>$45</td>
<td>$45</td>
<td>$4.66</td>
<td>$160</td>
<td>Linear Foot</td>
<td>12 (17)</td>
</tr>
<tr>
<td>Infrastructure</td>
<td>Description</td>
<td>Median</td>
<td>Average</td>
<td>Minimum</td>
<td>Maximum</td>
<td>Cost Unit</td>
<td>Number of Sources (Observations)</td>
</tr>
<tr>
<td>----------------</td>
<td>---------------------</td>
<td>--------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>-----------</td>
<td>---------------------------------</td>
</tr>
<tr>
<td>Sidewalk</td>
<td>Concrete Sidewalk + Curb</td>
<td>$170</td>
<td>$150</td>
<td>$23</td>
<td>$230</td>
<td>Linear Foot</td>
<td>4 (7)</td>
</tr>
<tr>
<td>Sidewalk</td>
<td>Sidewalk Unspecified</td>
<td>$34</td>
<td>$45</td>
<td>$14</td>
<td>$150</td>
<td>Linear Foot</td>
<td>17 (24)</td>
</tr>
<tr>
<td>Sidewalk</td>
<td>Sidewalk Pavers</td>
<td>$70</td>
<td>$80</td>
<td>$54</td>
<td>$200</td>
<td>Linear Foot</td>
<td>3 (4)</td>
</tr>
</tbody>
</table>
Chapter 6 - Conclusions

This study investigated the correlation between infrastructure deficiencies and pedestrian-vehicle and cyclist-vehicle crash frequency at intersections using ordered logistic regression and K-means clustering techniques. The models for this study were developed for a case study in Dallas, Texas, with 3,795 geocoded pedestrian and cyclist crashes at 33,965 intersections during the 2015-2019 period. The infrastructure deficiencies considered were sidewalk, crosswalk, and pavement deficiencies near intersections.

The results show that for intersections in low-income areas, the odds of having more pedestrian and cyclist crashes are 22% and 34% higher than intersections in medium-income and high-income areas, respectively. For intersections with sidewalk, crosswalk, or pavement deficiencies, the odds of having more pedestrian and cyclist crashes are 86%, 15%, and 29% higher than intersections without these deficiencies, respectively. For intersections with one, two, or three infrastructure deficiencies, the odds of having more pedestrian and cyclist crashes are 2.8, 3.0, and 3.2 times higher than intersections without infrastructure deficiencies, respectively. Overall, these findings indicate that the infrastructure deficiencies considered in this study are strongly correlated with crashes in high traffic areas, particularly in low-income neighborhoods that are more likely to have higher pedestrian and bicycle activity.

The intersections in Dallas were also categorized into four different groups based on their crash frequency and infrastructure deficiency level. Further investigation into each category indicates that those intersections with infrastructure deficiencies and crashes are associated with high annual average daily traffic (AADT), bus stop ridership, and pedestrian trips. These findings
indicate that intersections with these characteristics in low-income areas will likely incur the greatest benefits in reducing crash risk through infrastructure investments.

There are several limitations of this study, and more research is needed to further advance the work. First, the datasets used in this study may have changed during the period of the crash data (2015-2019), and such changes have not been reflected in the analysis. Second, the data used in this study were all from Dallas, Texas, and additional data from other cities would be helpful for generalizing the model and enhancing its applicability to more diverse conditions. Third, other infrastructures such as traffic signals, traffic signs, and bike lanes were excluded from the analysis due to incomplete spatial coverage. The same methodology could be applied with these features added to extend the analysis. Finally, the unit of analysis for this study is intersections. However, disability and vulnerability data were only available at the Census tract level, and car ownership and income data were only available at Census block group levels. If finer-scale data could be obtained, the models could be re-fit, potentially with different results.
Appendix

3.2.1 Intersection Dataset

The intersection dataset is downloaded from the City of Dallas (Dallas EGIS, 2021). This 2018 dataset includes 33,965 points, where each point indicates the intersection of two sections or streets or roads in Dallas. Figure 16 shows all of the intersection points within Dallas that are included in the study.

Figure 16 - Dallas intersections
3.2.2 Pedestrian and Cyclist Crash Dataset

The crash dataset is downloaded from Crash Records Information System (CRIS, 2021), which includes 3795 pedestrian and cyclist crashes from 2015 to 2019 in Dallas. Unfortunately, 614 of the crashes do not have latitude and longitude information; therefore, those crashes have been eliminated from this analysis. Figure 17 shows the locations of the pedestrian and cyclist crashes in Dallas.

Figure 17 - Pedestrian and cyclist crashes in Dallas
3.2.3 Median Household Income by Block Group Dataset

The 2019 5-year estimates of median household income by block group data are downloaded from the United States Census Bureau (Census, 2021). These data are inflation-adjusted based on U.S. dollars. The income level dataset has 52 missing values. The missing values are estimated using the average of the most recent available 5-year inflation-adjusted median household income of all block groups. Figure 18 shows the results of block groups classification based on income.

Figure 18 - The income level of block groups in Dallas
3.2.4 Pedestrian Trip Generation Dataset

The pedestrian trip generation data are downloaded from the ITE Trip Generation Web-based App (ITE, 2020). This app estimates trip generation for a specific location by travel mode and land use. For instance, Figure 19 shows that for a liquor store on a weekday during the afternoon peak hour, an average of 33 pedestrian trips are generated per 1000 sq. ft. gross floor area with a 21.7 standard deviation. The trip statistics are estimated by ITE personnel using a model based on data gathered at study sites in the United States. For instance, Figure 19 shows that the liquor store model is:

\[ T = 11.77(X) + 66.80 \]  

(8)

Figure 19 - The ITE pedestrian trip generation estimation for liquor store

![Graph showing pedestrian trip generation estimation for liquor store](image-url)
3.2.5 Bus Stop Ridership Data Dataset

The bus stop ridership data are downloaded from Dallas Area Rapid Transit (DART, 2021). The DART data include bus stop ridership data for each station in Dallas from October 2019 to February 2020 in three categories: weekday, Sunday, and Saturday. The data include the total number of people that have used a specific bus station in a specific month. For instance, in October 2019, 2,121 people used the Park Lane station in Dallas. Figure 20 shows the locations of the bus stops included in this study.

Figure 20 - Bus stops in Dallas
3.2.6 Annual Average Daily Traffic Dataset

The annual average daily traffic data are downloaded from the Texas Department of Transportation dataset (TxDOT, 2021). This dataset includes the annual average daily traffic from 2013 to 2018 for different segments of Dallas and has 16 missing values. Figure 21 shows the annual average daily traffic for different segments/streets in Dallas that are included in the study, as well as those that are missing.

Figure 21 - Annual average daily traffic for different roads/streets in Dallas
3.2.7 Vulnerability and Disability Dataset

The 2019 5-year estimates of vulnerability and disability by tract-level dataset are downloaded from United States Census Bureau (Census, 2021). Figure 22 shows the total population with disability in Dallas at the tract level, while Figure 23 indicates the total population of vulnerability in Dallas at the tract level.

Figure 22 - Total disability distribution in Dallas
Figure 23 - Total vulnerability distribution in Dallas
3.2.8 Lane Width Dataset

The lane width data are downloaded from the City of Dallas (Dallas EGIS, 2021). This dataset includes the lane width of different segments/roads in Dallas in 2018 and has 149 missing values. Figure 24 shows the lane width points of different routes and streets in Dallas, included in the study, as well as those that are missing.

Figure 24 - The lane width points of streets in Dallas
3.2.9 No Car Ownership Dataset

The car ownership dataset is downloaded from the United States Census Bureau (Census, 2021). This dataset includes the total number of tenures (privately owned lands or households) that do not have access to their own cars within Dallas Census tracts. Figure 25 shows the no-car-ownership percentage by tract level in Dallas.

Figure 25 - No car ownership distribution in Dallas
3.2.10 Land Use Dataset

The land use dataset is downloaded from Dallas Enterprise GIS (Dallas EGIS, 2021). This dataset contains the human use of each plot in Dallas in 2019 that can be categorized as residential, medical, industrial, educational, industrial, commercial, or recreational. Figure 26 summarizes how the land use data are categorized into six major groups, along with subsets of each group. Figure 26 shows all of the land use points in Dallas.

Figure 26 - Land use groups and subsets
3.2.11 Sidewalk Dataset

The sidewalk dataset was obtained from the Dallas Department of Public Works (DPW, 2021). The dataset classifies sidewalk conditions in 2019 into four categories: none, damaged, leave-out (partially missing), and damaged and leave-out. Figure 27 shows an example of damaged & leave-out sidewalk or sidewalk deficiency. Figure 28 indicates the locations in Dallas with sidewalk deficiencies.

Figure 27 - A damaged & leave-out sidewalk condition
3.2.12 Crosswalk Dataset

The crosswalk data are generated by Zheng (2021) using a deep learning object recognition algorithm called YoLo3 (Redmon, J., & Farhadi, A, 2018) and high-resolution Google satellite imagery in 2019. The dataset classifies crosswalks conditions at intersections into deficient or non-deficient categories. Crosswalk deficiency is defined as at least one missing or damaged crosswalk at intersections with traffic lights or stop signs within residential areas. Figure 29 shows an example of a damaged crosswalk. Figure 30 indicates the locations in Dallas with crosswalk deficiencies.
Figure 29 - An example of a crosswalk deficiency

Figure 30 - Locations with crosswalk deficiencies in Dallas
3.2.13 Pavement Dataset

The pavement dataset was obtained from the Dallas Department of Public Works (DPW, 2021). The dataset classifies pavement condition into a numerical index between 0 and 100, called pavement condition index (PCI). The PCI index is developed by the United States Army Corps of Engineers and is standardized by the American Society for Testing and Materials (ASTM). Deficiency for pavement condition means that the pavement has a PCI index below or equal to 55 (ARA, 2016). Figure 31 shows an example of pavement with PCI less than 55. Figure 32 indicates the locations in Dallas with pavement deficiency.

Figure 31 - A damaged & leave-out pavement condition
3.2.14 Database Validation

After creating and cleaning the final database, a sample of 5,000 intersections was randomly chosen among the 30,409 intersections for quality checking purposes. Google Street View was used to virtually assess each intersection in the sample to check the infrastructure deficiency data accuracy. Among 5000 samples, only 37 observations (0.7%) were found with infrastructure deficiency errors. The 37 errors in the final database were corrected manually. Figure 33 shows an example of errors found in the database. In Figure 33, there is no sidewalk for pedestrians, and some crosswalks are missing. However, the final database does not indicate any infrastructure deficiency for this particular intersection.
Figure 33 - An example of errors found in the database
4.2.1 The Box Plot Comparison Among Clusters for Other Features

The box plots comparing pedestrian and cyclist crash and infrastructure deficiency score distribution among clusters are presented in Section 4.2 in the main text, while the box plot comparisons among clusters for other features are presented below.

Figure 34 - AADT distribution among different clusters

Anova, p < 2.2e-16
Figure 35 - Bus stop ridership distribution among different clusters

Figure 36 - No car ownership distribution among different clusters
Figure 37 - Pedestrian trip distribution among different clusters

Figure 38 - Vulnerability distribution among different clusters
Figure 39 - Disability distribution among different clusters


56